

Detecting the Age of Twitter Users

Benjamin Paul Chamberlain¹, Clive Humby², Marc Peter Deisenroth¹

Department of Computing, Imperial College London, London SW7 2AZ¹

Starcount Insights, 2 Riding House Street, London W1W 7FA²

Abstract

Twitter provides an extremely rich and open source of data for studying human behaviour at scale. It has been used to advance our understanding of social network structure, the viral flow of information and how new ideas develop. Enriching Twitter with demographic information would permit more precise science and better generalisation to the real world. The only demographic indicators associated with a Twitter account are the free text name, location and description fields. We show how the age of most Twitter accounts can be inferred with high accuracy using the structure of the social graph. Besides classical social science applications, there are obvious privacy and child protection implications to this discovery. Previous work on Twitter age detection has focussed on either user-name or linguistic features of tweets. A shortcoming of the user-name approach is that it requires real names (Twitter names are often false) and census data from each user's (unknown) birth country. Problems with linguistic approaches are that most Twitter users do not tweet (the median number of Tweets is 4) and a different model must be learnt for each language. To address these issues, we devise a language-independent methodology for determining the age of Twitter users from data that is native to the Twitter ecosystem. Roughly 150,000 Twitter users specify an age in their free text description field. We generalize this to the entire Twitter network by showing that age can be predicted based on what/whom they follow. We adopt a Bayesian classification paradigm, which offers a consistent framework for handling uncertainty in our data, e.g., inaccurate age descriptions or spurious edges in the graph. Working within this paradigm we have successfully applied age detection to 700 million Twitter accounts with an F1 Score of 0.86.

Introduction

Owing to its popularity and openness, Twitter data is frequently used in scientific research. In its raw form it contains limited information about the people generating the data. Enriching Twitter data with age estimates is useful for scientific modelling and has commercial applications in areas such as marketing and personalised advertising.

Kosinski, Stillwell, and Graepel (2013) have shown that people's interests are indicative of their age and that Facebook page likes can be used to predict age with high accuracy. Page likes play a similar role in Facebook as follows

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

do in Twitter and so this result indicates that the same results could be achieved with Twitter. There are two principle causes of this effect: Firstly, people of similar ages tend to have similar interests as a result of age related life events (education, child birth, marriage, employment, retirement, wealth changes etc.). Secondly, social networks exhibit homophily with people tending to connect with others having similar attributes (age in this case), and because information about what else to follow flows through connected people (McPherson, Smith-Lovin, and Cook, 2001).

Previous research into Twitter age detection has focussed on linguistic models of tweets or the free text "user name" field. Linguistic models require significant research efforts to be applied for every distinct language and has currently only been developed in Dutch (Nguyen et al., 2013). In addition, any method that relies on tweets is unable to make predictions for the majority of Twitter users who do not generate tweets in adequate volumes. This is a problem since the median number of tweets in our data set of 700 million users is four. Methods based on user-name rely on knowledge of the relative frequency of the user's true name in their country of birth (Oktay, Firat, and Ertem, 2014). However, Twitter users are under no obligation to supply their real names and their country of birth is generally unknown.

In this paper, we take a different approach based on the hypothesis that (a) someone's interests depend on their age, and (b) what they follow describes (some of) their interests. We use this hypothesis to derive a machine learning model that predicts a user's age from what/whom they follow. We take the 133,000 accounts of users who specify their age in their description field as labelled ground truth. The features of the model are the 103,722 accounts that have greater than 10,000 followers and are followed by more than 10 labelled accounts. The features are subsequently used to infer the age of any Twitter user based on whom/what they follow within a Bayesian framework. Unlike methods that rely on linguistic style or first-name frequencies, our method generalises across national and linguistic boundaries. However, a major challenge with this data is that ages specified in the description field can be inaccurate and the graph contains many spurious edges that are not indicative of true interests. For this reason, we adopt a Bayesian classification paradigm, which offers a consistent framework for handling all forms of uncertainty.

Related Work

One of the first studies of Twitter demographics was conducted by Mislove, Lehmann and Ahn (2011). They used a large corpus of Tweets to derive demographic data for Twitter users located in the US. They did not study age, choosing instead to focus on the distributions of geography, race and gender. They were able to conclude that the Twitter population is not representative of the broader US populace.

Previous work on Twitter age detection has focussed on either aspects of language or user names. For instance, Nguyen et al. (2013) produced a lexical analysis of Dutch language tweets. They obtained ground truth through a manual tagging process employing several students. The principle features were simply unigrams, where the intuition was that older people use more positive language, fewer pronouns and longer words and sentences. Other features like intensifiers (emoticons), alphabetic lengthening (e.g., “sweeeeeet”) and capitalisation were also useful. They found that age prediction works well for young people, but that above the age of 30, language tends to homogenise. They also noted a strong relationship between language and gender (women tend to identify more with language and use more first singular pronouns, while men use more links) implying that they need to be modelled together. They used simple linear models in the study (linear and logistic regression). Age detection using lexical features is complicated by the existence of the concept of social age (Meder, 2011). A social age is determined by life stage (married, children, employment etc.) rather than years since birth and it has a strong affect on writing style. Social age and gender are fluid constructs and attempts to predict them from text oversimplify the problem. For instance, people interacting in groups of the opposite sex may adapt their language to mimic the perceived social norm.

Okta, Firat, and Ertem (2014) estimate the age of Twitter users from the first name supplied in the free-text account-name field. They use US Social Security data to generate probability distributions for birth year given age and show that for some names age distributions are quite sharply peaked. They also used second names to estimate ethnicity and found that both age and ethnicity have large effects on the patterns of Twitter usage.

Kosinski, Stillwell, and Graepel (2013) used Facebook-likes to determine a broad range of attributes of users. They generated ground truth through responses to a Facebook app designed to measure a user’s personality. They were particularly successful at identifying age (80% accuracy on a held out test set), and their approach has the benefit of generalising readily to different locales. Facebook likes are closely related to Twitter follows; both are indicators of user interest. However their work differs from ours for two principle reasons. Firstly Facebook contains a birthday field and so age can easily be calculated given access to this field. Secondly, they assumed that users accessing their app represented a stratified sample of Facebook ages and we are not able to make this assumption.

Finally we make use of a survey of internet using Americans conducted by Duggan and Brenner (2013). They identified a sample of 1802 adults (speaking either English or

Table 1: Age categories and counts. Mean features gives the average number of feature accounts followed.

idx	age range	count	freq	mean features
0	under 12	7753	5.9%	23.7
1	12–13	20851	15.8%	27.9
2	14–15	30570	23.1%	30.8
3	16–17	23982	18.1%	28.7
4	18–24	33331	25.2%	26.0
5	25–34	9286	7.0%	23.1
6	35–44	3046	2.3%	22.6
7	45–54	1838	1.0%	16.0
8	55–64	962	0.7%	11.4
9	over 65	596	0.5%	11.2

Spanish) using random cold calling and recorded their demographic information and use of social media. 288 of their respondents were Twitter users.

Data and Preliminaries

Only roughly 0.02% of Twitter profile descriptions contain age data and so working with a small sample of the network is not viable. We implemented a distributed web crawler to download the Twitter graph and associated account metadata using Twitter access tokens mined through an app. To optimize data throughput while remaining within Twitter’s rate limits we developed an asynchronous system connected to an access token server using Python’s Twisted library (Wysocki and Zabierowski, 2011). We indexed each user’s description field using Apache SOLR and searched the index for REGular EXpression (REGEX) patterns that were indicative of age data across Twitter’s four major languages (English, Spanish, French and Portuguese). Initially, we attempted to match the multilingual equivalents of the English phrases, such as “I am x ”, “I am x y”, “ x years”. This resulted in significant issues with phrases of the form “more than x years”, “think I am x years” and “feel like x years”, which required further refinements of the REGEX matching string to logically exclude these forms, e.g., “ x years NOT(feel like x years, think I am x years)”. This process discovered 133,000 users who disclosed their age (i.e., 0.02 % of the 700 million indexed accounts) with a total of 15.9 million outgoing edges of which 7.14 million were to accounts with greater than 10,000 followers.

We bucket ages of users into ten categories chosen to suit our use case of targeted marketing. The age categories and number of accounts, relative frequency and average number of features per category are shown in Table 1.

Table 1 compares the relative counts of age classes reported in user descriptions to the counts that would be expected based on scaling up the survey conducted by Duggan and Brenner (2013). Our data set appears to be highly skewed towards younger Twitter users. We believe this is because older users are less inclined to disclose their age. As a result, any model that attempts to use this sample to make broader predictions must take this skew into account.

Table 1 also shows that data is relatively small for ages

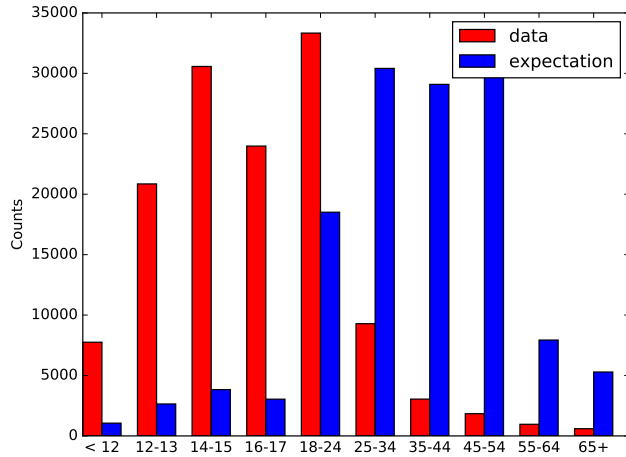


Figure 1: Counts of ages within each age category. The data is heavily skewed towards younger users when compared with an expectation based on user surveys (Duggan and Brenner, 2013).

greater than 45, an effect that is compounded by older users tending to follow fewer features. To increase the data set size for improving the predictive performance for these age categories, we looked for terms that implicitly define older users. Using occupation terms to infer age bands induced new problematic biases (principally levels of education and affluence). However, we discovered that while older Twitter users do not disclose their age, they disclose that they are retired or that they are parents or grandparents in large numbers. To identify parents or grandparents it is necessary to exclude second person mentions (e.g., “my grandmother”), and we found that any mention of both a male and female parent or grandparent tended to be erroneous. We assumed that retired people were over 65 years old and so it was necessary to exclude the large numbers of retired military personnel who are now working in civilian jobs. To ensure mutual exclusivity we allocated accounts that indicated multiple age bands to the most specific, assuming that “retired” is more specific than “grandparent”, which in turn is more specific than “parent”. Table 2 shows the number of users who describe themselves as parents, grandparents or retired and the age categories that we assumed for them. Counts are orders of magnitude greater than those for users who described themselves having a specific age with higher average features followed that for any class in Table 3.¹

Age Detection based on Follows

Given a set of 133,000 labelled data points we wish to predict the age of the remaining 700 million Twitter users within our data set. The standard machine learning approach to achieve this task is to measure a set of features that define

¹We did not use parents in our work due to their low age specificity. We include the data as a sign post to future researchers who may wish to use parents as a starting point to efficiently sample older Twitter users for manual age coding.

Table 2: Twitter users who disclose that they are parents, grandparents or retired in their descriptions. Data is selected to be mutually exclusive, i.e., “parents” is “parents” AND NOT (“grandparents” OR “retired”) and “grandparents” is “grandparents” AND NOT “retired”. We did not calculate the average number of features exhibited by parents.

	index	age range	count	mean features
parents		over 18	1,485,164	-
grandparents		over 45	63,895	45.0
retired		over 65	176,748	36.1

a space in which the labelled data can be separated, and then to learn specific parameters for each class.

Feature Selection

Our hypothesis is that someone’s age is predictive of what they are interested in and that Twitter-follows serve as a proxy for interests (similar to Facebook-likes). For this reason we selected Twitter accounts with greater than 10,000 followers as the features of our model. We only used features that are followed by more than 10 users who disclosed their age giving 103,722 features. Table 3 shows the 10 features with the highest support (number of followers who disclose their age) in our labelled data. The followers column gives the total number of followers of each feature across the Twitter network. There is a Pearson correlation of 0.86 between the support and the total follower count for our data set.

Bayesian Model

We adopt the Bayesian modelling paradigm because it provides the only way to consistently model the various types of uncertainty (noisy labels, bad edges, survey estimates etc.) encountered in this problem. In what follows we denote the feature vector of an unlabelled account by $X \in \{0, 1\}^M$, where M is the number of features and $X_i = 1$ if and only if the user follows the i^{th} feature, with the corresponding target vector of ages as $A \in \{0, 1, \dots, 9\}$. We use $\mathbf{X} \in \{0, 1\}^{N \times M}$ to represent the whole labelled data set with N the number of labelled data points and $\mathbf{A} \in \{0, 1, \dots, 9\}^N$ the known labels.

We wish to estimate $P(A|X, \mathbf{X}, \mathbf{A})$, the distribution of the age A of an unseen user X given the set of accounts \mathbf{X} that they follow and the labelled training set \mathbf{A} . As older people are less likely to explicitly state their age, our labelled data is severely biased and we can not simply assign multinomial distributions $P(A|X_i)$ for each feature based on the relative frequencies of their followers (see Table 3). To overcome this problem we make use of Bayes’ law in the form

$$P(A|X, \mathbf{X}, \mathbf{A}) \propto P(X|A, \mathbf{X}, \mathbf{A})P(A)$$

and learn the probability of following each feature account given an age category. Here $P(A)$ is the prior distribution of Twitter user ages, which we take from the survey by Duggan

Table 3: The 10 accounts with the highest support within the labeled data set. Followers gives the total followers across all of Twitter

twitter_handle	support	under 12	12-13	14-15	16-17	18-24	25-34	35-44	45-54	55-64	65+	followers
justinbieber	20359	1517	5179	5737	4202	3073	412	99	67	34	38	5.6×10^7
katyperry	18395	1467	4180	4410	3604	3575	701	158	124	75	102	5.9×10^7
taylorswift13	15199	1207	3417	3674	3045	2919	507	113	117	79	122	4.6×10^7
selenagomez	14264	1270	3578	3691	2847	2339	367	76	43	26	27	2.4×10^7
ArianaGrande	13512	1254	3404	3604	2631	2172	319	50	40	19	20	2.0×10^7
ddlovato	13259	1099	3284	3562	2741	2135	301	53	37	19	28	2.5×10^7
onedirection	12834	979	3472	3778	2767	1622	138	43	20	7	8	2.1×10^7
Harry_Styles	12830	912	3468	3936	2751	1581	120	24	15	9	13	2.2×10^7
NiallOfficial	12498	858	3431	3895	2702	1468	90	24	15	8	8	2.0×10^7
YouTube	11688	926	2496	2687	2193	2287	495	183	154	99	169	4.6×10^7

and Brenner (2013)²:

$$P(A) = [0.8, 2, 22, 3, 14, 23, 23, 22, 6, 4] \times 10^{-2}$$

Our model has 10^5 features, which implies 5×10^9 pairwise correlations. For tractability we make the Naive Bayes' assumption that the decision to follow each account is independent given the age of the user, leading to

$$P(X|A, \mathbf{X}, \mathbf{A}) = \prod_{X_i=1} P(X_i|A, \mathbf{A}, \mathbf{X})$$

where i indexes the features. We only consider cases where $X_i = 1$ as the Twitter graph is very sparse³, and the default position of every account is to follow nobody. Therefore, the act of not following an account does not contain enough information to justify the additional computational cost. The Maximum Likelihood Estimator (MLE) is given by

$$P(X_i = 1|A, \mathbf{A}, \mathbf{X}) = \frac{n_{i,a}}{n_a}$$

where n_a is the number of labelled data points in age category a and $n_{i,a}$ is the number that also follow account i .

After applying Laplacian smoothing with parameter equal to 1 (Bishop, 2006) to $P(X_i = 1|A, \mathbf{A}, \mathbf{X})$ the point estimate model produces many pathological distributions with high probability in incorrect categories or bi-modal class predictions of the form "under 12 with 50% probability and over 65 with 50% probability". The weakness of this model is that it assumes the same level of uncertainty for each $P(X_i = 1|A, \mathbf{A}, \mathbf{X})$ when in reality the distributions for Twitter's most popular accounts are less uncertain than those with only a few thousand followers. To resolve these issues we model $P(X_i|A, \mathbf{A}, \mathbf{X}, \mu)$ as a Bernoulli random variable and place a conjugate Beta prior on the Bernoulli parameter μ . Figure 2 shows the corresponding probabilistic graphical model that we use to infer the age of Twitter users. Probabilistic graphical models (PGMs) are tools used to visualise independencies inherent in a collection of random variables and are described in great detail in Koller and

Friedman (2009). In Figure 2 the labelled training data is represented as $\mathbf{A}_{i,j}$ and $\mathbf{X}_{i,j}$ and the target value to be predicted as A . Given the learnt model parameters $\mu_{i,a}$, A only depends on the set of features X_i . Figure 2 describes the following generative process for age data:

1. For each feature/age combination draw $\mu_{a,i} \sim \text{Beta}(b, c)$
2. Choose an age $A \sim \text{Mult}(\pi)$
3. For each feature $X_i \sim \text{Bernoulli}(\mu_{i,a})$

Due to conjugacy, the posterior distribution is also Beta distributed. By integrating out μ in the usual Bayesian way we obtain

$$P(X_i|A, \mathbf{X}, \mathbf{A}) = \int_0^1 P(X_i|\mu_i, A) P(\mu_i|b_i, c_i) d\mu_i \quad (1)$$

$$= \int_0^1 \mu_i P(\mu_i|b_i, c_i) d\mu_i \quad (2)$$

$$= \frac{n_{i,a} + b_i}{n_a + b_i + c_i} \quad (3)$$

We seek hyper-parameters b, c of the prior $p(\mu)$ that smooth out noisy observations of less popular feature accounts, but which do not have a large effect when ample data is available. To achieve this we set c_i to be constant across all features (hence dropping the subscript) and proportional to the total number of observations in each age class. We then set b_i so that $\mathbb{E}[\mu_i] = \frac{n_i}{N} = \frac{b_i}{b_i + c}$. This generates a prior that assumes that the a-priori probability of following an account is constant across age classes and varies in proportion to the number of followers across features.

Data Cleaning

Applying REGEX matches to free-text fields inevitably leads to some false matches owing to unanticipated character combinations when working with large data sets. In addition, many Twitter accounts, while correctly labelled, may not represent the interests of human beings. This can occur when accounts are controlled by machines (bots), accounts are set up to look authentic to distribute spam (spam accounts) or account passwords are hacked in order to sell authentic looking followers. To detect and remove spurious labelled data points we looked for accounts that had

²The survey only looked only looked at over-18s and so we inferred under-18s based on US census data

³We measured 700 million nodes having 50 billion edges, which implies a density of 1.6×10^{-7}

under 12	135	151	148	43	99	69	25	2	0	0
12-13	164	324	661	176	291	133	45	1	0	0
14-15	110	259	1039	453	552	192	63	1	0	0
16-17	46	111	624	484	587	210	63	0	0	0
18-24	68	72	355	429	1028	631	168	1	0	0
25-34	14	14	35	41	229	346	95	2	0	0
35-44	4	4	8	5	40	125	68	0	0	0
45-54	5	2	4	4	25	58	41	1	0	0
55-64	2	1	0	2	10	34	25	1	0	0
65+	1	0	4	8	7	13	13	1	0	0
	under 12	12-13	14-15	16-17	18-24	25-34	35-44	45-54	55-64	65+

(a) Confusion matrix for a 10 % held out test set. Data included only accounts with explicit age labels. Very few predictions are made in older categories where labelled data is limited.

under 12	125	156	135	42	95	30	12	35	5	16
12-13	145	343	613	191	254	71	25	60	5	24
14-15	109	266	1005	478	462	134	43	103	32	46
16-17	51	120	567	465	503	118	41	101	25	45
18-24	71	83	307	461	880	341	89	280	70	88
25-34	13	22	44	49	167	193	50	150	39	49
35-44	4	2	11	8	36	47	41	59	13	9
45-54	7	14	28	35	79	72	68	1610	2475	670
55-64	8	15	24	33	64	63	53	2110	2093	682
65+	25	34	91	163	342	196	198	6936	6451	6051
	under 12	12-13	14-15	16-17	18-24	25-34	35-44	45-54	55-64	65+

(b) Confusion matrix using retired people and grandparents.

Figure 3: Confusion matrices for a 10 % held out test set. (a) Data included only accounts with explicit age labels. Very few predictions are made in older categories where labelled data is limited. (b) By including data from retired people and grandparents, the predictions in the older age categories are improved substantially.

a large impact on the likelihood $P(\mathbf{X}|\mathbf{A})$. The Kullback-Leibler divergence (KL divergence) is the most commonly used to compare two probability distributions. We calculated an anomalousness score for the i^{th} data point

$$S_i = KL(P||P_i), \quad (4)$$

where P is the likelihood calculated from the full, labelled data set and P_i is the likelihood calculated from the labelled data set minus the i^{th} data. We excluded any training examples that were more than three Median Absolute Deviations (MAD) from the mean score. This process excluded 246 accounts from our training data. Examples of invalid training data are shown in Table 4.

Experimental Evaluation

We report the performance of our ten category age model using the directly labelled age data described in Table 3. For comparison with Nguyen et al. (2013) we also report F1 scores, precision and recall for a three-category model with categories under 18, 18–45 and over 45. In this second set of experiments, we make use of the data labelled as grandparents and retirees described in Table 2. In all experiments, we trained our model using 90% of the data and evaluated performance using the remaining 10 % of the data.

In Table 5, we report the five features with the highest posterior age values of $P(A|X_i = 1)$ for each age category a of A^4 . The account descriptions are taken from the first line

⁴To aid with description generation we restricted the table en-

of the relevant Wikipedia page. The youngest Twitter users are characterised by an interest in internet celebrities and computer games players. Genres of music are important in differentiating all age groups from 12–45. 25–34 year olds are in part marked by entities that saw greater prominence in the past. This group is also distinguished by an interest in pornographic actors. Age categories 45–54 and 55–64 have the same top five and are differentiated by their interest in religious topics. Users over the age of 65 are identifiable through an interest in certain (retirement) sports and politics.

Figure 3a shows a confusion matrix for classification using just the data of Table 3. In the case of perfect predictions, the confusion matrix is diagonal. We can identify the diagonal structure in Figure 3a, but the model based on this data set has huge difficulties to make predictions for ages greater than 45, and classifies these accounts largely in the range 25–44. This is because higher age classes have low counts and mean features (See Table 3) and so many features have no support in these age classes. This is supported by the corresponding performance statistics for each class, which are presented in Table 6, together with aggregates. The table shows that the predictive performance rapidly drops off for ages greater than 35.

Figure 4 shows a learning curve used to estimate the extent our model could be improved by adding additional data. The horizontal axis shows the amount of data used for training to English language accounts with more than 100,000 followers

Table 4: Spurious data points identified by taking the Median Absolute Deviation of the hold one out KL-Divergence.

Handle	Description	REGEX age	Notes
RIAMOpera	Opera at the Royal Irish ... Presenting: Ormindo Jan 11...	11	An Irish Opera
TiaKeough13	My name Tia I'm 13 years old .	13	Hacked account
39yearoldvirgin	I'm 39 years old... if you're a woman, I want to meet you.	39	Probably not 39
50Plushealths	Retired insurance Agent After 40 years of Services.	retired	Using reciprocation software
MrKRudd	Former PM of Australia... Proud granddad of Josie & McLean...	grandparent	Outlier. Former AUS PM

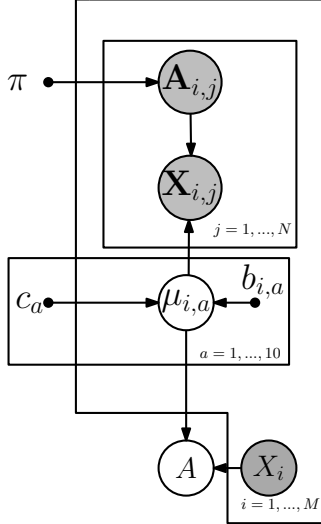


Figure 2: The probabilistic graphical model for the age estimation problem. Circles are random variables, with shading when they are observed. Small dots indicate model parameters and the plates denote replication.

ing. The remaining data is used to test the model. The large gap between the classification performance of the training data and the held out test data, combined with the upward trajectory of the test classification curve, indicates that our model would benefit from additional data.

To improve the predictive performance we added the data for grandparents and retired people described in Table 2. Retired people were all added to the 65+ category. Grandparents were randomly allocated to either 45–54, 55–64 or 65+. Figure 3b shows the confusion matrix using the enhanced data set. Predictive performance is greatly improved in the higher age categories, and the weight around the main diagonal of the matrix is improved for all categories.

Table 7 shows the performance statistics for this augmented data set. F1 scores for the top three age classes are dramatically improved in comparison to Table 6. As a result the macro F1 is 25 % higher, however the micro-F1 score is only marginally improved due to small degradations in other categories.

For comparison with the state-of-the-art work of Nguyen et al. (2013) based on linguistic we consider the performance of our model as a three-class classifier using the following age bands: under 18, 18–44 and 45+. This choice allows the allocation of grandparents to a single class and has class

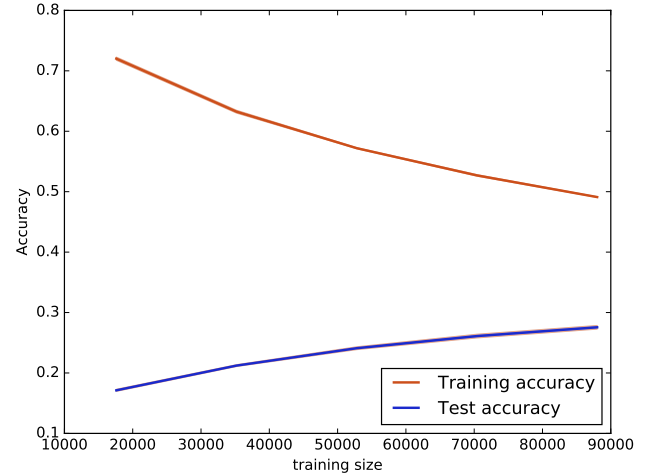


Figure 4: Learning curves showing the classification accuracy on the training and test set as the size of the training set is increased for a three-class model. The chart indicates that predictive accuracy on unseen data could benefit from additional training data. Error bars are one standard deviation on 5-fold cross-validation

boundaries that align with the ten-class model.

Figure 5 shows the three-class confusion matrix. The mode for each row is on the main diagonal and the main cause of error is a misclassification of accounts in class 18–45 as under 18.

Table 8: Performance statistics for a three-class age model.

metric	under 18	18–44	over 45
support	7096	3676	30690
recall	0.68	0.50	0.95
precision	0.76	0.39	0.96
F1	0.72	0.44	0.96
micro F1	0.86		
macro F1	0.58		

The corresponding performance statistics are shown in Table 8 are very good with a micro F1 score of 0.86; the same value was achieved by Nguyen et al. (2013). However, the major advantage of our model is that we have applied it to the entire Twitter graph, as opposed to being limited to a biased sample of Dutch Twitter users with a sufficient

Table 5: Most discriminative features for all age categories.

twitter_handle	description	under 12	12–13	14–15	16–17	18–24	25–34	35–44	45–54	55–64	65+
Under 12-year olds											
RosannaPansino	vlogger	0.40	0.22	0.15	0.09	0.07	0.02	0.01	0.01	0.01	0.02
AntVenom	minecraft gamer	0.40	0.25	0.15	0.09	0.06	0.02	0.01	0.01	0.01	0.01
Bajan_Canadian	internet personality	0.37	0.25	0.17	0.10	0.06	0.02	0.00	0.01	0.01	0.01
shaycarl	vlogger	0.36	0.20	0.14	0.10	0.07	0.04	0.02	0.02	0.02	0.02
InTheLittleWood	gaming commentator	0.34	0.23	0.16	0.11	0.08	0.02	0.01	0.01	0.01	0.02
12–13 year olds											
ivandorschner	child TV presenter	0.18	0.27	0.20	0.11	0.09	0.03	0.03	0.02	0.03	0.04
Vikkstar123	youtuber	0.29	0.26	0.20	0.14	0.07	0.02	0.01	0.01	0.01	0.02
PeytonList	child actress	0.29	0.25	0.20	0.14	0.07	0.02	0.01	0.01	0.01	0.01
G_Hannelius	child actress	0.31	0.25	0.18	0.13	0.07	0.02	0.02	0.01	0.01	0.01
Cimorelliband	girlband	0.20	0.25	0.23	0.17	0.09	0.02	0.01	0.01	0.01	0.01
14–15 year olds											
therealsavannah	child pop singer	0.10	0.18	0.27	0.21	0.12	0.02	0.01	0.03	0.03	0.03
jessicajarrell	child pop singer	0.12	0.21	0.26	0.24	0.10	0.02	0.01	0.01	0.01	0.01
TheDylanHolland	child R&B singer	0.12	0.22	0.26	0.24	0.11	0.02	0.01	0.01	0.01	0.01
OfficialBirdy	child singer	0.10	0.17	0.26	0.24	0.13	0.04	0.01	0.02	0.02	0.02
officialjman	child singer	0.10	0.18	0.26	0.28	0.13	0.02	0.01	0.01	0.01	0.01
16–17 year olds											
TannerPatrick	singer	0.05	0.13	0.25	0.30	0.18	0.03	0.01	0.01	0.01	0.01
TheWordAlive	metalcore band	0.04	0.11	0.19	0.29	0.22	0.09	0.02	0.01	0.01	0.01
MitchLuckerSS	deathcore singer	0.05	0.14	0.23	0.29	0.20	0.04	0.01	0.01	0.01	0.02
metrostation	electronic band	0.03	0.07	0.15	0.29	0.18	0.10	0.04	0.06	0.06	0.03
BreatheCarolina	electronic band	0.06	0.15	0.22	0.29	0.19	0.06	0.01	0.01	0.01	0.01
18–24 year olds											
wecameasromans	metalcore band	0.05	0.13	0.22	0.28	0.21	0.06	0.01	0.01	0.01	0.01
Sum41	rock band	0.07	0.11	0.18	0.24	0.21	0.09	0.02	0.02	0.03	0.03
hopsin	rapper	0.04	0.09	0.13	0.19	0.20	0.09	0.09	0.06	0.05	0.07
Diablo	computer game	0.03	0.06	0.09	0.13	0.20	0.17	0.09	0.05	0.06	0.12
paparoach	rock band	0.04	0.09	0.14	0.19	0.20	0.12	0.07	0.06	0.06	0.04
25–34 year olds											
icp	hip hop duo	0.02	0.04	0.05	0.09	0.19	0.37	0.09	0.04	0.05	0.05
kevinrichardson	boyband member	0.02	0.03	0.05	0.09	0.16	0.35	0.12	0.07	0.06	0.04
skulleeroz	boyband member	0.02	0.04	0.06	0.09	0.16	0.33	0.12	0.07	0.06	0.05
LeeEvansNews	comedian	0.02	0.03	0.06	0.07	0.17	0.32	0.09	0.08	0.09	0.09
miko.lee	adult actress	0.04	0.03	0.03	0.05	0.17	0.31	0.08	0.07	0.08	0.14
35–44 year olds											
djspooky	hip hop artist	0.01	0.02	0.03	0.02	0.04	0.15	0.45	0.14	0.06	0.08
Mr_Mike_Jones	rapper	0.01	0.01	0.01	0.01	0.03	0.14	0.44	0.16	0.09	0.10
HISTORYTV18	history TV channel	0.02	0.03	0.03	0.05	0.09	0.14	0.36	0.10	0.06	0.13
TopDawgEnt	record label	0.03	0.07	0.05	0.07	0.11	0.11	0.36	0.09	0.03	0.07
DannySwift	boxer	0.02	0.03	0.04	0.07	0.06	0.09	0.33	0.12	0.08	0.16
45–54 and 55–64-year olds (identical most-discriminant features)											
JohnBevere	evangelist	0.00	0.00	0.00	0.00	0.01	0.01	0.07	0.36	0.39	0.15
edstetzer	evangelist	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.36	0.39	0.16
ChristineCaine	evangelist	0.00	0.00	0.01	0.00	0.01	0.01	0.07	0.36	0.38	0.15
womenoffaith	faith group	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.36	0.38	0.16
RELEVANT	faith magazine	0.00	0.01	0.00	0.01	0.01	0.01	0.07	0.35	0.38	0.17
People over 65											
afneil	political journalist	0.00	0.00	0.01	0.01	0.02	0.02	0.04	0.17	0.25	0.48
Chris_Boardman	retired cyclist	0.01	0.01	0.01	0.02	0.01	0.01	0.04	0.17	0.25	0.47
SkySportsGolf	golf TV channel	0.01	0.02	0.02	0.02	0.03	0.01	0.04	0.16	0.22	0.46
IamAustinHealey	retired rugby player	0.04	0.02	0.01	0.01	0.01	0.01	0.04	0.17	0.25	0.45
anthonyfjoshua	boxer	0.02	0.03	0.03	0.04	0.09	0.06	0.03	0.08	0.15	0.45

Table 6: Statistics for age prediction using only explicitly labelled data.

metric	under 12	12–13	14–15	16–17	18–24	25–34	35–44	45–54	55–64	65+
support	672	1795	2669	2125	2752	776	254	140	75	47
recall	0.20	0.18	0.39	0.23	0.37	0.45	0.27	0.01	0	0
precision	0.25	0.35	0.36	0.29	0.36	0.19	0.11	0.10	0	0
F1	0.22	0.24	0.38	0.26	0.37	0.27	0.16	0.01	0	0
micro F1	0.30									
macro F1	0.21									

Table 7: Statistics for age prediction using explicitly labelled data and people described as retired or grandparents.

metric	under 12	12–13	14–15	16–17	18–24	25–34	35–44	45–54	55–64	65+
support	651	1731	2678	2036	2670	776	230	5058	5145	20487
recall	0.19	0.20	0.38	0.23	0.33	0.25	0.18	0.32	0.41	0.30
precision	0.22	0.33	0.36	0.24	0.31	0.15	0.07	0.14	0.19	0.79
F1	0.21	0.27	0.37	0.24	0.32	0.19	0.10	0.20	0.26	0.43
micro F1	0.31									
macro F1	0.25									

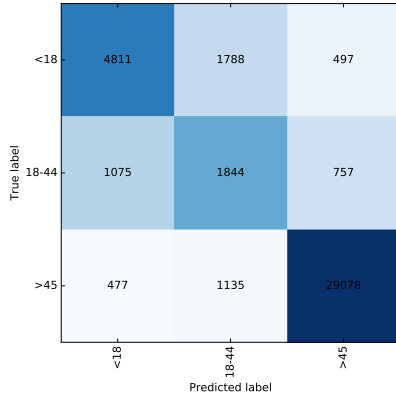


Figure 5: Three-class confusion matrix using all explicitly labelled data and people described as grandparents of retired.

number of Tweets.

Figure 6 shows the areas under the receiver-operator characteristics (ROC) curves for the three-class model. The curves are generated by measuring the true positive and false positive rates for each class for a range of classification thresholds. A perfect classifier has an area under the curve (AUC) equal to one, while a completely random classifier follows the dashed line with an $AUC = 0.5$. Performance is excellent for classes under 18 and over 45, but weaker for 18–45 where training data was more limited. The three-class learning curve (Figure 7) indicates that performance could still be improved by adding more labelled data. A promising way to do this would be to manually label accounts self-identifying as parents, however we lacked the resources to achieve this at meaningful scale.

We have evaluated the performance of a 10-class age

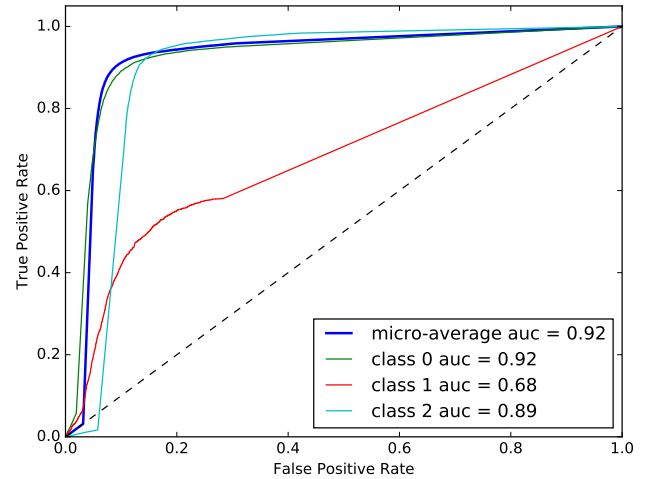


Figure 6: Receiver operator characteristics for three class age detection (0 = under 18, 1 = 18–45, 2 = 45+). The dashed line indicates random performance.

model trained on Twitter description data, which performs well with implicitly labelled age data. Our model performs as well as the current state of the art on a three-class problem for inferring the age of Twitter users without being limited to specific linguistic features. Finally, we have successfully applied our model to the entire Twitter network. Figure 8 shows aggregate classification results for 700 million Twitter accounts compared with expectations based on survey data (Duggan and Brenner, 2013). The figure shows that our model predicts that over 50% of Twitter users are between 18 and 35. Much of the bias of the original training set (Figure 1) has been removed, largely due to Bayesian inference.

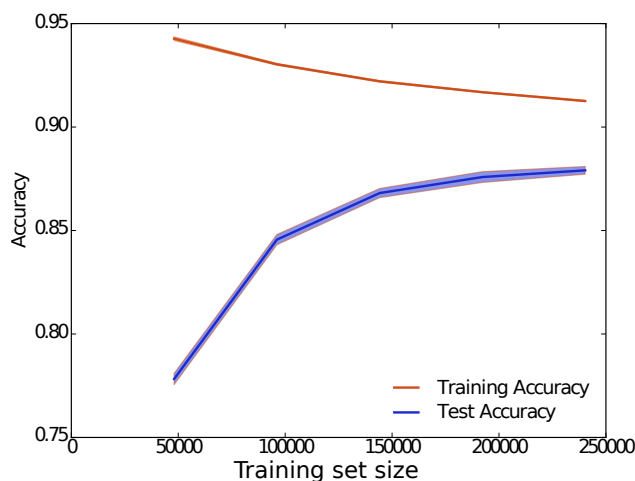


Figure 7: Learning curves showing the classification accuracy on the training and test set as the size of the training set is increased for a three class model. Error bars are one standard deviation on 5-fold cross-validation

Conclusion and Future Work

We have presented the first model that can classify the age of most Twitter users. Our work allows future researchers to build more complete models of social processes as measured on Twitter and to generalise phenomena on Twitter to the real world. The key idea behind our model is that a user's age can be predicted from what/whom they follow. Our Bayesian model delivers state-of-the-art performance equal to that achieved by the linguistic model of Nguyen et al. (2013), but with much broader applicability and more general assumptions. We are able to achieve these results using only the biased training data that is available natively within Twitter and so do not rely on manual labellers or an external application to produce training data.

Extensions include the combination of our results with linguistic models to attain higher fidelity estimates for particular locales and relaxing the independence assumptions between features. It is computationally intractable to explicitly model all dependencies, but algorithms, such as probabilistic PCA, are able to discover the most significant correlations efficiently Tipping and Bishop (1999).

Acknowledgements

This work was partly funded through a Royal Commission for the Exhibition of 1851 Industrial Fellowship.

References

- Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer Verlag.
- Duggan, M., and Brenner, J. 2013. *The Demographics of Social Media Users-2012*.
- Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

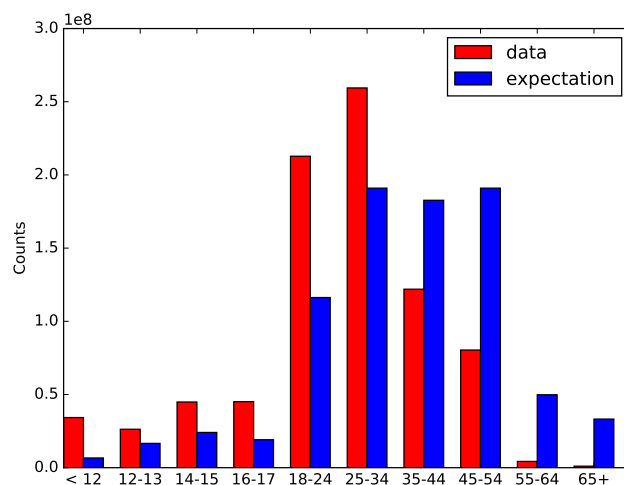


Figure 8: Predicted age proportions across the whole of Twitter compared to expectations based on survey data.

- Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America* 110(15):5802–5.
- McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444.
- Meder, T. 2011. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *International Conference on Computational Linguistics*.
- Mislove, A.; Lehmann, S.; Ahn, Y.Y. 2011. Understanding the Demographics of Twitter Users. In *International Conference on Computational Linguistics*.
- Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. “How old do you think I am?” A study of language and age in Twitter. In *International Conference on Weblogs and Social Media*, 439–448.
- Oktay, H.; Firat, A.; and Ertem, Z. 2014. Demographic breakdown of Twitter users: An analysis based on names. *Conference on Big Data, Socialcom, Cybersecurity*.
- Tipping, M., and Bishop, C. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*.
- Wysocki, R., and Zabierowski, W. 2011. Twisted framework on game server example. *International Conference on Experience of Designing and Application of CAD Systems in Microelectronics*.